公共数据资源目录的制作与管理

顾立平 [12]
1.中国科学院文献情报中心，北京，100190
2.中国科学院大学经济管理学院信息资源管理系，北京，100190

【摘要】本研究探索公共数据资源目录的制作和管理，以及数据出版和传播的重要性。公共数据资源目录的核心目标是促进公共数据的开放与利用，提高数据的价值。科学数据资源目录规划时需要重点考虑科学数据资源的类型、数据开放共享的权益管理，以及科研生命周期中产生数据和使用数据的场景等。数据出版和传播的重要性在于提高数据的开放利用价值和目标选择的有效性，促进科学数据的共享和再利用。数据集可通过开放许可协议对外提供使用，但使用者需遵守协议中的使用要求。在访问和共享方面，机构应当提供公共使用的数据文件、限制使用的数据文件以及时效性要求，旨在平衡数据的开放利用和保护产权之间的关系。
【关键词】公共数据资源目录；数据出版传播；开放许可协议；数据共享；产权保护。

Production and management of public data resource catalogs

GU Liping[12]
1.National Science Library, Chinese Academy of Sciences
2.Department of Information Resource Management, School of Economic and Management, University of Chinese Academy of Sciences

【Abstract】This study explores the production and management of public data resource catalogs, as well as the importance of data publication and dissemination. The core goal of the public data resource catalog is to promote the openness and utilization of public data and enhance its value. When planning the directory of scientific data resources, it is necessary to focus on the types of scientific data resources, the management of rights and interests for open and shared data, as well as the scenarios for generating and using data in the scientific research lifecycle. The importance of data publication and dissemination lies in improving the open utilization value and effectiveness of target selection of data, promoting the sharing and reuse of scientific data. Datasets can be provided for external use through open license agreements, but users are required to comply with the usage requirements in the agreement. In terms of access and sharing, institutions should provide publicly available data files, restricted data files, and timeliness requirements, aiming to balance the relationship between open utilization of data and protection of property rights.
【Keywords】Public data resource directory; Data publication and dissemination; Open License Agreement; Data sharing; Property protection.

一、公共数据资源目录的制作

数据资源目录的设立，其核心目标在于促进公共数据的开放与利用，进而提升数据的价值。通过目录，人们可以充分了解国家已经开放共享了哪些数据集和数据产品，这有助于社会各界及时接入国家数据共享交换平台，充分调动社会和相关部门在公共数据开放共享上的积极性，进一步提高数据利用程度。

目前受到广泛应用的数据分类标准，是美国国家科学委员会（National Science Board, NSB 2015）构建的三种类别标准:

（1）研究型数据集：某个项目或者多个项目的科研产出的结果之一，数据加工的程度较低，通常在一个研究团队里产生，并且数据集仅仅在团队成员内共享，项目结束后可能就不会保存这类数据，而有数据流失的风险。

（2）社区型数据集：根据已有的行业标准或者规范，人们建立这类数据集，在社区内进行交流。他们可能受到直接资助产生，除了社区或者资助来源的优先使用权利之外，往往没有明确持续维护它们的承诺或者举措。

（3）参考型数据集：这类数据集主要产生并且用于国际范围的科学社群，不仅数据集的成本预算较大，而且在学术社区内具有多样性和分布广的特点，这类数据集已有领域范围内的数据管理的方式，往往遵守健壮性标准，并且致力长期保存。

从科学数据管理的角度看，不同的科学生命周期阶段会产生不同类型的数据集，这些数据集具有不同的使用价值和含义。研究型数据集、社区型数据集和参考型数据集分别代表了科研的不同阶段和成果，每种类型的数据集都有其独特的价值和作用。

对于科研管理人员来说，考虑到数据的产生、规模、规范和共享等因素，需要有针对性地选择数据管理方式。研究型数据集更适合以数据知识库的方式进行存储和管理，社区型数据集则更适合在数据中心及其平台进行管理，而参考型数据则需要全球性学科化的数据共享平台或国家级学科数据中心进行管理。

科学数据资源的类型、数据开放共享的权益管理，以及科研生命周期中产生数据和使用数据的场景等，都是科学数据资源目录规划时需要重点考虑的内容。通过科学合理的规划和管理，可以进一步提高公共数据的开放利用价值和目标选择的有效性，为社会发展和科技进步提供更多有力的支持。

二、公共数据资源目录的管理

数据资源目录的使用对于科研人员来说具有重要的价值，它可以帮助他们有效地管理数据、统一引用标识符，并提高数据的可发现性，进而促进数据的长期保存。为了实现公共数据的开放利用价值，科学数据资源目录的设计需要考虑以下五个方面的措施：

（1）功能设计：科学数据资源目录应具备开放获取、数据重用和长期保存三大功能，以满足科研人员的需求。这些功能的规划设计应根据附属机构的要求和目标进行选择，以确保目录的实用性和有效性。

（2）收录方式设计：收录方式可以分为主动和被动两种，其中自动收割和科研人员、机构自主存储是两种常见的形式。选择何种收录方式应根据实际情况和目标考虑，以确保数据资源的全面性和及时性。

（3）元数据设计：元数据是确保科研数据可发现、可理解、可重用和可持续发展的基础。因此，完善的元数据设计至关重要。除了对数据内容的描述外，还应设置贡献者元数据和许可权的元数据，为数据创建者设定使用权限提供条件。

（4）提交规范：科学数据资源目录应对数据提交者的身份、提交数据的类型、提交时间和提交方式进行明确规定。这有助于保证目录数据的准确性和可靠性，同时为数据的更新和维护提供便利。
（5）数据管理与运营规划：资金政策和人才政策是数据管理政策的重要组成部分，对于数据的正常更新、保存和利用具有重要影响。因此，在制定数据管理与运营规划时，应充分考虑这些因素，以确保数据服务运营的稳定性和可持续性。

国外科学数据中心的管理规范值得我们借鉴。例如，地震科学注册研究中心（IRIS）的发展战略指出，提供计算资源的访问，使得科学社区更容易执行定制化的科学研究。这体现了公共数据开放利用的价值和目标选择的重要性，为科研社区提供了更加便捷和高效的数据服务[01]。

三、公共数据产品服务

数据发布对于公众审查科研人员研究成果和新知识的发现具有重要意义。为了推动科研数据的发布，可以制定严格的发布期限。然而，为了保障数据产品的有效发挥与利用，科研数据的发布也可以采取订阅付费机制或根据商业合同进行数据转移使用的接口服务。在选择科研数据许可权设置标准时，科研团队可以根据自身条件选择不同的许可标准。科研数据的许可权设置标准主要包括六种情况：知识共享许可 CC（Creative Commons licenses）[02]、公共领域许可 CC0\PDDL[03-04]、开放数据协议许可（ODC）[05]、开放政府许可（OGL）[06]、限制性许可证（RL）[07]和设计科学的许可（Design Science Licence, DSL）[08]等。人们可以采用复合许可方式，允许数据接受者选择一种方式来使用数据。

不同学科的数据类型、特征和发表规则不同，因此需要根据学科特点制定差异化的数据管理政策。例如，地震科学注册研究中心（IRIS）提供了数据质量估计和可视化功能，以成为科学社区中数据的“受信任源”。同时，对于统计数据，要求标明时间、地点、国家、区域码、数据类型等信息[09]。

国际科学数据中心通过数据分析和挖掘，可以形成有价值的科学数据产品，并开展相关增值服务。这可以鼓励社会组织和企业开展市场化增值服务，同时也可以促进科研人员最大化开发利用科学数据共享的潜在价值。通过公共数据的开放利用，可以推动科学研究的进步和创新，为社会发展和科技进步提供更多有力的支持

四、数据出版和传播

科技期刊在数据共享和数据出版方面扮演着重要的角色，它们一直致力于将科学成果的再现作为高质量学术刊物对科学界所应担负的责任。近十年来，科技出版界在科学论文和科学数据的引用以及科研人员在这方面的贡献方面取得了长足的进步。许多国际期刊都制定了科学数据共享政策，要求作者在投稿时提供相关科学数据或提供可获得这些数据的第三方数据存储库的存取号，以促进数据的共享和再利用。比如： Biodiversity Data Journal[10]、Ecology[11]、Earth System Science Data[12]等。这些数据期刊也遵循 GDPP 的数据保护规范，对作者进行数据管理的规范。

科学数据中心为提高科学数据的传播工作，多支持科研人员整理发表产权清晰、准确完整、共享价值高的科学数据。例如，地震科学注册研究中心（IRIS）的数据服务政策要求任何团体或组织单位在传播其数据时，需明确数据来源并要求数据接收者确认数据来自 IRIS，以保障数据的可追溯性和权威性[13]。这样的

措施有助于促进公共数据的开放利用，提高数据的价值和目标选择的有效性，为科学研究和社会发展提供更多有力的支持。

在国际学术出版业中，随着科技期刊数据共享政策的推进，数据期刊和数据论文应运而生。以自然出版集团（NPG）为例，其期刊出版政策制定了促进科学数据共享的严格政策，要求作者必须无条件地提供材料、数据和协议给他人使用[14]。在此基础上，2014 年 5 月，NPG 参考了科研资助机构的数据要求，以及科研人员、图书馆员、数据知识库管理员和数据标准倡议者对科研数据管理方式的调查结果、相关权益和意见，推出了在线出版的开放获取期刊 Scientific Data。该期刊以论文的类型发布具有科学价值的数据描述，即数据描述[15-16]。

SD 中的数据集主要包括计算或策划数据以及通过实验或观察产生的数据，其中包括“技术验证”和“用法说明”部分[17]。这些数据描述符可用于描述已在其他出版物中分析到的数据集，或用于描述独立的数据集，遵循一定的数据标准，计算机可读、可检索。这些数据描述符并不包括新的科学方法的描述或新的科学假说的测试。

为确保数据的质量和开放性，不同的学科类别会成立专门的编辑委员会来对数据进行辨识判定。大多数数据描述符将接受至少一个具有相关实验技术专业知识的科学家和一个数据标准专家的评审。在审核过程中，需要考虑实验方法的有效性、第三方使用数据的完整性、数据描述符与数据内容的连贯性，以及数据能否被开放获取和使用等因素。这些措施有助于提高公共数据的开放利用价值和目标选择的有效性，为科学研究和社会发展提供更多有力的支持。

五、数据引用规范

数据引用不仅涉及文献之间的参考文献，还包括数据与文献、数据集与数据集、数据与数据之间的多重关系。对数据引用的标示是对数据贡献者成果的认可和尊重，也是保障数据提供者权益的重要基础。具体包括但不限于以下四大类：

（1）学科/数据/机构知识库的 ID 识别码：引用数据时，建议使用知识库授予科研数据的 ID 识别码。不同知识库会根据数据的属性，采用不同的引用方式。如机构知识库中，德国科学技术信息服务机构、大英信息服务机构、澳大利亚国家数据服务中心(Australian National Data Service，ANDS)等机构利用由 DataCite 专门为科研数据集分配可作为独立的、可引用的永久标示符；学科知识库中，如生命科学领域的寡核苷酸多态性数据库（database of SNP, 简称 dbSNP）则采用由 SNP 的 ss 号和 refSNP 号的数据库标识符（简称 RS 号）[18]。数据知识库蛋白质组数据库建议在引用的同时，指出 PX 标识符，使数据集更可见和可访问[19]。

（2）学科/数据/机构知识库网址：除了引用 ID 识别码外，也可以直接引用存储科研数据的知识库网址。这种方式有助于增加数据的可见性和可访问性。如 ArrayExpress 功能基因组学实验数据库要求在引用数据时包括数据的识别符和 ArrayExpress 主页网址（www.ebi.ac.uk/arrayexpress）[20]。

（3）数据论文和原文标示符：一些知识库会推荐采用 DataCite 的引用格式或类似格式，包括引用数据的识别符和相应的原文标示符。如 GEO（Gene Expression Omnibus）建议提交者引用其识别符（GSExxx），同时也建议用户引用他人的原文和该文章所对应数据记录的识别符[21]。

（4）数据提供者所要求引用的学术论文：在某些情况下，数据提供者会规定需要引用的学术论文。在引用过程中，应按照要求予以标示。

通过不同的引用方式，可以提高公共数据的开放利用价值和目标选择的有效性，促进科学数据的共享和再利用，为科学研究和社会发展提供更多有力的支持。

在数据引用的技术线路和实施方式上，主要依赖于关联开放数据和知识元库两大途径。

（1）通过关联开放数据，可以跨越不同的数据源，执行复杂查询，将开放信息转换为可计算的开放知识[22-24]。这有助于提高数据的可重用性和结构化程度，同时增强数据的语义化和关联化。关联开放数据的过程包括一系列处理程序和规则，如采集数据、公布关联数据、规范化链接中的"连接点"等，以促进数据的开放获取和共享利用。

（2）知识元库是由开发者编译、分发和维护的大型数据库，提供有关电子资源的信息。它为解决适当副本问题而设计，支持论文论点的证据索引，包括实验数据、图表等[25-29]。知识元库的供应链由多个角色组成，如出版商、内容持有者、订阅代理者、图书馆和信息中心等。通过制定标准，可以解决数据质量和许可证问题，促进数据的规范化和共享利用[30-32]。

这些技术线路和实施方式有助于提高公共数据的开放利用价值和目标选择的有效性，为科学研究和社会发展提供更多有力的支持。通过数据的关联开放和知识元库的应用，可以促进数据的共享、再利用和创新，推动科学进步和社会发展。

六、公共数据的无偿提供与收费标准

在明确数据产权的基础上，数据集可以通过开放许可协议对外提供使用，但使用者需遵守协议中的使用要求。开放许可授权是在拥有知识产权和使用限制的前提下进行的。例如，美国校际社会科学数据共享联盟（ICPSR）在访问和共享的方面[33]，声明 ICPSR 会把项目的科学数据提供给的社会科学研究界使用，实施方式主要具有三项要点：（1）公共使用的数据文件：这些文件中的直接和间接标识符信息已被删除，以减少泄露风险，可以通过 ICPSR 网站对其直接访问。同意使用条款后，拥有 ICPSR 账户和其成员单位授权 IP 地址的用户可以下载数据，非会员用户可购买数据。（2）限制使用的数据文件：这些文件分布在这些数据中，即除去数据的潜在识别信息将显著损害数据的分析潜力。用户（和其所在机构）必须对这些文件提出访问申请，创建数据安全计划，并同意其他访问控制协议。（3）时效性：项目中的研究数据需在项目结束之前提交给 ICPSR，以便围绕数据的可用性所产生的任何问题都能够得到解决。延迟发布数据是可能的。延迟发布政策允许将数据在双方共同商定的期限内存储起来而不对外发布（通常是一年）。公共使用的数据文件可直接访问和下载；限制使用的数据文件需申请访问并遵守其他协议；研究数据需在项目结束前提交给 ICPSR。这些措施旨在平衡数据的开放利用和保护产权之间的关系。

对于经过处理加工的数据集，需要根据智力劳动程度和资产性质进行界定和明确。这些数据产品可能由企业、特定科研机构或设备制作完成，制作方和委托方之间的合同会规定产权归属和使用条件。因此，在对外收费原则上，应遵循合同中约定的具体内容进行数据产品的收费标准与价格设定。

这样的开放利用模式可以保护数据产权，同时促进数据的共享和利用，提高公共数据的价值。通过合理的许可协议和收费标准，可以鼓励更多的人使用和研究这些数据，推动科学和社会的发展。

七、结语

公共数据中心的共享与利用机制对于公共数据的开放利用具有重要的价值和目标选择。通过制作和管理科学数据资源目录，可以促进数据的共享和流通，提高数据的利用效率和效益。同时，科学数据产品的服务和出版传播也可以促进科研成果的审查和新知识的发现，推动科技进步和社会发展。

在制作和管理科学数据资源目录方面，我们需要充分调动社会和部门在科学数据开放共享上的积极性，推动数据的共享和流通。通过设计合理的功能、收录方式、元数据和提交规范，可以提高数据资源目录的质量和可用性，方便用户了解和使用已开放共享的数据集和数据产品。

在数据出版和传播方面，科技期刊在数据共享和数据出版方面已经取得了长足的进步。许多期刊都制定了科学数据共享政策，推动了数据的开放获取和共享。这有助于促进科研成果的审查和新知识的发现，提高科研的质量和效率。

数据引用规范也是公共数据开放利用的重要方面。通过制定学科/数据/机构知识库的 ID 识别码、文献引用、数据集引用和数据引用等规范，可以促进数据的可追溯性和可重复性，提高数据的质量和可信度。同时，开放许可协议也可以促进数据的开放获取和利用，提高数据的利用效率和效益。

公共数据的开放利用具有重要的价值和目标选择。通过科学数据中心的共享与利用机制，我们可以促进数据的共享和流通，提高数据的利用效率和效益，推动科技进步和社会发展。我们应该重视公共数据开放利用的价值，选择合适的目标来实现其最大化的效益。

参考文献

[01] Incorporated Research Institutions for Seismology (IRIS), Data Services. Data Services [EB/OL].[2023-08-16].https://www.iris.edu/hq/files/programs/data_services/policies/Strategic_Plan_v7.pdf
[02] Creative Commons.Share your work[EB/OL].[2023-08-16]https://creativecommons.org/
[03] CC0[EB/OL].[2023-09-22].http://creativecommons.org/publicdomain/zero/1.0/
[04] PDDL[EB/OL].[2023-09-22].http://opendatacommons.org/licenses/pddl/
[05] ODC Attribution-SharealikeCommunityNorms[EB/OL].[2023-09-22].http://opendatacommons.org/norms/odc-by-sa/
[06] Open Government Licenceforpublic sector information[EB/OL].[2023-09-22]. http://www.nationalarchives.gov.uk/doc/open-government-licence
[07] AusGOAL Restrictive Licencetemplate[EB/OL].[2023-09-22].http://www.ausgoal.gov.au/restrictive-licence-template
[08] DSL [EB/OL].[2023-09-22].http://www .gnu.org/licenses/dsl.html
[09] Incorporated Research Institutions for Seismology (IRIS), Data Services. Data Services [EB/OL].[2023-08-16].https://www.iris.edu/hq/files/programs/data_services/policies/Strategic_Plan_v7.pdf

[10] Biodiversity Data Journal. For Author [EB/OL]. [2023-08-16]. https://bdj.pensoft.net/
[11] Journal of Ecology. Guides to Better Science [EB/OL]. [2023-08-16]. https://www.britishecologicalsociety.org/publications/guides-to/
[12] Earth System Science Data. Data protection agreement [EB/OL]. [2023-08-16]. https://www.copernicus.org/data_protection.html /.
[13] Incorporated Research Institutions for Seismology (IRIS), Data Services. IRIS Data Services Policy Regarding Redistribution of IRIS Data Policy Version 2.0 [EB/OL].[2023-08-16].https://www.iris.edu/hq/files/programs/data_services/policies/Redistribution_Policy.V2.0.pdf
[14] Nature Publishing Group. Availability of Data and Materials[EB/OL]. [2023-08-16]. http://www.nature.com/authors/policies/availability. html..
[15] Nature Publishing Group.Data publication survey-raw data[EB/OL]. [2023-08-16]. http://figshare.com/articles/Data_publication_survey_raw_data/1234052.
[16] Scientific Data[EB/OL]. [2023-08-16]. ..
[17] Scientific Data. Format of Data Descriptors.[EB/OL].[2023-08-16].http://www.nature.com/sdata/for-authors.
[18] NCBI. dbSNP [EB/OL].[2023-08-16]http://www.ncbi.nlm.nih.gov/snp
[19] EBI. PRIDE Archive [EB/OL].[2023-08-16] http://www.ebi.ac.uk/pride/archive/
[20] ArrayExpress. Submitting data to ArrayExpress (general) [EB/OL].[2023-08-16]. http://www.ebi.ac.uk/arrayexpress/help/faq.html#cite.
[21] GEO. Citing and linking to the GEO database [EB/OL] [2023-08-16]http://www.ncbi.nlm.nih.gov/geo/info/linking.html.
[22] Gorlitz, O., Staab, S.. Federated Data Management and Query Optimization for Linked Open Data[J]. New Directions in Web Data Management, 2011, 331:109-137.
[23] Omitola T., Koumenides CL., Popov IO., et al.. Put in Your Postcode, Out Comes the Data: A Case Study[C]. 7th Extended Semantic Web Conference (ESWC2010), Semantic Web: Research and Applications Proceedings, 2010, 6088:318-332.
[24] Schomburg S.. Publishing Aleph data as Linked Open Data[EB/OL] [2023-09-22] .
[25] Van de Sompel, H., Hochstenbach, P.. Reference Linking in a Hybrid Library Environment. Part 3: Generalizing the SFX Solution in the "SFX@Ghent & SFX@LANL" experiment[J/OL]. D-Lib Magazine, 1999, 5(10): http://dx.doi.org/10.1045/october99-van_de_sompel
[26] Van de Sompel, H., Hochstenbach, P.. Reference Linking in a Hybrid Library Environment. Part 2: SFX, a Generic Linking Solution[J/OL]. D-Lib Magazine, 1999, 5(4): http://dx.doi.org/10.1045/april99-van_de_sompel-pt2.
[27] Van de Sompel, H., Hochstenbach, P.. Reference Linking in a Hybrid Library Environment. Part 1: Frameworks for Linking[J]. D-Lib Magazine, 1999, 5(4):http://dx.doi.org/10.1045/april99-van_de_sompel-pt1.
[28] Van de Sompel H., Beit-Arie O.. Open Linking in the Scholarly Information Environment Using OpenURL Framework [J].D-Lib Magazine, 2001, 7(3):www.dlib.org/ dlib/march01/vandesompel/03vandesompel.html.

[29] Beit-Arie O.. Linking to the Appropriate Copy: Report of a DOI-Based Prototype[J]. D-Lib Magazine, 2001, 7(9):http://www.dlib.org/dlib/september01/caplan/09caplan.html.
[30] Chandler A.. NISO IOTA: Improving OpenURLs Through Analytics, in Context[EB/OL].[2023-08-16]
[31] Culling J.. Link Resolvers and the Serials Supply Chain[EB/OL][2023-08-16] http://www.uksg.org/projects/linkfinal.
[32] Jewell T., Aipperspach J., Anderson I., et al. Making Good on the Promise of ERM: A Standards and Best Practices Discussion Paper[EB/OL][2023-08-16] .
[33] ICPSR. Data Management Plans[EB/OL].[2023-03-08]http://www.icpsr.umich.edu/files/datamanagement/DataMa